Vers un Procédé d'extraction automatisé de Connaissances

Joseph Sobel Ndiaye

QUANTICS F

Recherche & Développement en Sciences de l'Information

Mars 2021

La problématique de la surcharge informationnelle n'est point de l'ordre du « Big Data » , mais une question de « Bad Data ».

- Réciproquement, exclure du « Bad Data » consiste à extraire du « **Good Data** ».
- Le « **Good Data** » est beaucoup plus que de la donnée, plus que de l'information ; c'est de la connaissance.
- La Connaissance est une information efficiente, orientée vers un sujet et qui permet la prise de décision.
- L'extraction de connaissances, en raison du caractère non absolu que revêt pareille démarche, ne peut consister en une recherche d'optimum.
- L'extraction de connaissances à partir d'un corpus de texte en language naturel est donc une recherche de **Satisfactum**.
- Le procédé extractif en question diffère du résumé de texte.
- Le résumé automatisé de texte (principalement extractif) s'effectue sur des critères statistiques relatifs à la distribution des mots, leurs positions dans le corpus ; ou se base sur des modèles de type probabiliste.
- La limite de ces approches de type « arithmétique » demeure dans le fait que les dimensions sémantiques voire cognitives ne sont pas valorisées.
- **Illustration** :

  Nous pouvons extraire de la connaissance à partir d'une unité lexicale minimale : Un (**1**) mot.

	Résumé de texte	Extraction de connaissances
Phrase(s)	Possible	Possible
1 Phrase	Difficile	Possible
1 Mot	Impossible	Possible

➡ Exemple : À partir du mot ***Conscience***, ayant **36** synonymes ; nous avons identifié les mots ***advertance***, ***esprit***, ***minutie***, ***moralité***, comme ayant une **plus haute contribution à la connaissance du Concept** induit par le mot Conscience.

✳ La **hiérarchisation sémantique par critère de validité informative** permet d'envisager une automatisation de l'extraction de connaissances, à partir d'un corpus de texte en language naturel.